COMPARATIVE GENOMICS AT THE VERTEBRATE EXTREMES

Dario Boffelli, Marcelo A. Nobrega and Edward M. Rubin

Annotators of the human genome are increasingly exploiting comparisons with genomes at both the distal and proximal evolutionary edges of the vertebrate tree. Despite the sequence similarity between primates, comparisons among members of this clade are beginning to identify primate- as well as human-specific functional elements. At the distal evolutionary extreme, comparing the human genome to that of non-mammal vertebrates such as fish has proved to be a powerful filter to prioritize sequences that most probably have significant functional activity in all vertebrates.

In Homer's classic, Odysseus is confronted with the impossible task of steering his ship midway through a narrow strait, flanked by two rocks. The slightest veer to one side and he and his crew would be within reach of the monster Scylla, with its six teeth-filled horrible heads that can destroy anything within their reach. Swerve to the other side, and they would now fall prey to Charybdis, a whirlpool that sucks anything that comes close enough down to the bottom of the ocean[1].

Odysseus' quandary over the route to take evokes a contemporary dilemma that biologists face when they attempt to use comparative genomics to identify functional elements, such as genes, gene regulatory elements and other less well-defined structural components of the genome, buried in otherwise anonymous sequences. If they compare species that are too closely related, then the high degree of similarity between the orthologous sequences will obscure the functional elements within them; by contrast, if they compare species that are too distantly related, then the functional elements will have diverged too much to be readily identifiable[2]. Until recently, it was assumed that comparing species that are separated by moderate evolutionary times, such as non-primate mammals, would represent an ideal strategy[3,4].

The principle of 'steering a middle course' has certainly proved successful: comparative analyses of the human and mouse genomes, which diverged from each other approximately 75 million years ago, allowed much better annotation of both these genomes than would have been possible had only 1 been available[5,6]. Moreover, recent comparisons of multiple moderately related species have been extremely powerful in the analysis of simple genomes such as that of yeast. By comparing the genomes of four related *Saccharomyces* species, Kellis *et al.* not only revised the total count of yeast genes, but more importantly, identified most of the transcription-factor binding sites (TFBSs) in the yeast genome[7]. A similar approach applied to complex mammalian genomes indicates that multiple species comparisons might be a more efficient approach to the identification of putative functional elements than human–mouse comparisons alone[8] (BOX 1).

Two important issues, however, limit the use of this 'middle of the strait' approach to identifying functional elements in the human genome. First, some functional elements are certainly either human- or primate-specific, and accordingly, will be missing from the genomes of non-primates[9]. Second, the enormous degree of non-coding sequence conservation that is found between humans and mice is probably the consequence of non-uniform rates of evolution across the human genome[10]. This results in genomic sequences that still show a considerable degree of similarity that reflects a slower evolutionary rate rather than purifying selection[11,12].

The comparisons of the human genome sequence with those of extremely close (primates) and extremely distant (non-mammalian vertebrates) species have recently been demonstrated as an alternative to overcome these limitations, providing important new insights. The evolutionary distance between the species that are chosen for sequence comparison largely determines what kind

DOE Joint Genome Institute, Walnut Creek, California 94598, USA, and Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA. Correspondence to E.M.R. e-mail: emrubin@lbl.gov

doi:10.1038/nrg1350

Box 1 | Comparisons of multiple, moderately related species

Increasing the number of species that are used in genome comparisons makes it progressively less probable that sequences are conserved by chance, and helps in the identification of truly functionally conserved sequences to be prioritized for experimental analysis. Conserved sequences can be identified using statistical approaches that are designed to distinguish genomic regions that are evolving significantly more slowly than the local rate of neutral evolution[7,15,74]. Researchers from the National Institutes of Health (NIH) Comparative Sequencing Program applied these methods to the analysis of a 1.8-Mb interval from several mammalian species and identified many conserved non-coding sequences that were not identified by human–mouse comparisons alone. Conserved sequences that are identified by multiple species comparisons ultimately need to be studied experimentally to confirm their functional role. Frazer *et al.* recently tested a subset of conserved sequences that were identified by multiple mammal comparisons and found that they were significantly more likely to drive reporter gene expression and to bind to nuclear proteins than non-conserved sequences[75].

of functional elements can be identified. Only the most constrained functional elements can be identified by comparisons with distant, non-mammalian vertebrates. Conversely, the study of primate-specific biological functions is limited to comparisons that use primate genomes.

Choosing species to be used in comparative genomics represents a compromise, with benefits and limitations that need to be recognized and weighed. Understanding this principle, Odysseus ordered his unsuspecting crew to steer the boat close to Scylla, knowing that at least six hapless men would meet with an awful fate, but preventing the entire crew from perishing in the clutches of Charybdis. Fortunately, modern day biologists are not presented with Odysseus' dilemma. They are no longer offered only a single path, but rather, several paths to choose from, stemming from the sequencing of an increasing number of genomes. Here, we focus on comparative genomics at the phylogenetic extremes, using vertebrate sequences that are both closely related to and distant from *Homo sapiens* as an alternative form of comparative genomics. We highlight the sorts of insight that are offered by these strategies, their potential and limitations, and outline our vision of where the field of comparative genomics might be moving in the near future.

Distant species comparisons

One of the chief observations that immediately stemmed from the sequencing of the mouse genome was the unexpectedly large amount of conservation between humans and mice: 40% of the human and mouse genomes could be aligned at the nucleotide level[12,13]. On the other hand, only approximately 5% of the human genome (~150 Mb) seems to have evolved more slowly than the NEUTRAL RATE, with more than half of these sequences corresponding to non-coding sequences. A few of these non-coding sequences — identified in the portion of the human genome that could be aligned to the mouse genome — have no detectable activity in functional assays[14]. This indicates that the available assays might be inadequate and/or that many of these elements are indeed non-functional and are only conserved across mammalian genomes owing to an asymmetric rate of neutral evolution[15].

So, how are biologists going to sift through this sea of sequence conservation and prioritize those sequences that are most amenable to functional analysis? One obvious strategy would be to use species that are more distant from humans than mice, such as non-mammalian vertebrates, to identify a subset of sequences that are conserved over greater evolutionary distances. A leading visionary in the field, S. Brenner, proposed more than a decade ago that the compact genome of the teleost pufferfish *Fugu rubripes* was ideally suited as an innovative resource for deciphering the human genome[16]. Given the extreme phylogenetic separation between fish and mammals, it was reasonably assumed that only important functional sequences would be conserved between genomes that are otherwise so diverged. Brenner's prophecy was convincingly demonstrated with the completion of the sequencing of the *F. rubripes* genome, after which its first comparison to humans immediately revealed more than 1,000 genes that had previously been unidentified in the human genome[17].

Although initial interest in *F. rubripes* focused on its use as a gene-identification resource, the concept that human–fish comparisons would be useful for identifying *cis*-regulatory elements was not emphasized. These elements are generally plastic owing to their modular structure[18], which allows individual components, such as TFBSs, to evolve independently. Nevertheless, Aparicio *et al.* identified regulatory sequences in the vicinity of *Hoxb4* using mouse–*F. rubripes* comparisons[19]. Several recent studies have confirmed that human–fish comparisons can efficiently facilitate the identification of functional non-coding sequences.

One excellent example of how useful such comparisons can be is the use of human–*F. rubripes* comparisons to reveal multiple regulatory elements around *DACH*, a gene involved in embryonic development that has a complex expression pattern (FIG. 1). In mammals, 2 large GENE DESERTS surround *DACH*, so it is the only gene in a 2630-kb segment of the genome. Human–mouse comparisons of this interval revealed, in addition to conserved exons, more than 1,000 conserved non-coding sequences with >100 bp and 70% identity (an arbitrary criterion of conservation that is frequently used as a reasonable empirical significance threshold in many biological studies). This large number of non-coding elements makes it impractical to test individual conserved sequences for biological activity. So, a comparison between human sequence and that of several distant vertebrates, including a frog and 3 fish, was carried out, reducing the number of conserved non-coding

NEUTRAL RATE
Genetic variation that does not affect the fitness of the organism is not subject to selection and evolves at the neutral rate.

GENE DESERTS
Gene-poor regions in the genome that are larger than 500 kb. Gene deserts often contain sporadic evidence of transcription.

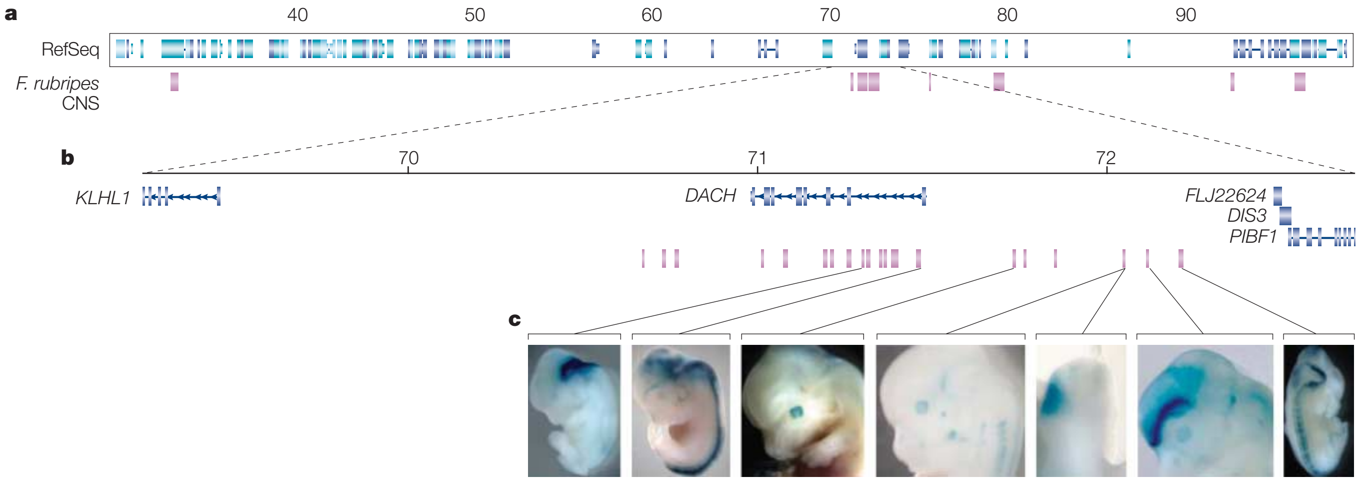


Figure 1 | **Architecture of human–*Fugu rubripes* conserved non-coding sequences in the human genome. a** | A 65-Mb segment of human chromosome 13 is shown that contains 145 well-characterized RefSeq genes (exons in blue). There are 51 human–*F. rubripes* conserved non-coding sequences (CNSs) in this region, which are distributed non-uniformly in clusters that contain 1–32 CNSs each (in purple). **b** | One cluster of human–*F. rubripes* CNS is illustrated in more detail. *DACH* — the only human gene in this region — is involved in key aspects of embryonic development. **c** | Testing some of the non-coding sequences that are conserved in humans and *F. rubripes* revealed that several of these elements correspond to enhancers in mouse embryos. In this assay, the sequence being tested is cloned upstream of a β-galactosidase reporter gene. If the cloned sequence is an enhancer, it will activate the reporter gene, which can be detected in an assay that stains the tissues that express β-galactosidase (in blue). Adapted with permission from REF. 20. © (2003) M. A. Nobrega and E. M. Rubin.

sequences to 32. An *in vivo* mouse transgenic assay on nine of these elements showed that seven of them were enhancers recapitulating several aspects of the endogenous expression of *DACH*[20]. These results illustrate that gene deserts can harbour sequence elements with crucially important biological functions, and support the idea that *cis*-regulatory sequences can affect gene expression at near-megabase distances[21–23]. A surprising characteristic of these 7 enhancers is their astonishing, near-absolute degree of conservation between humans and rodents: each enhancer contained a block of ungapped aligned sequence ranging from 250 to 530 bp with a degree of identity ranging from 98% to 99.5% between humans, mice and rats (BOX 2).

Several other studies have also shown that non-coding sequences that are conserved between humans and fish frequently correspond to elements with enhancer activity[24–31]. Although this does not exclude the possibility that a large fraction of the sequences that are conserved only among mammals are also enhancers, it does highlight the practical use of distant species comparisons, in that they maximize the likelihood of choosing non-coding sequences for analysis with a measurable biological activity.

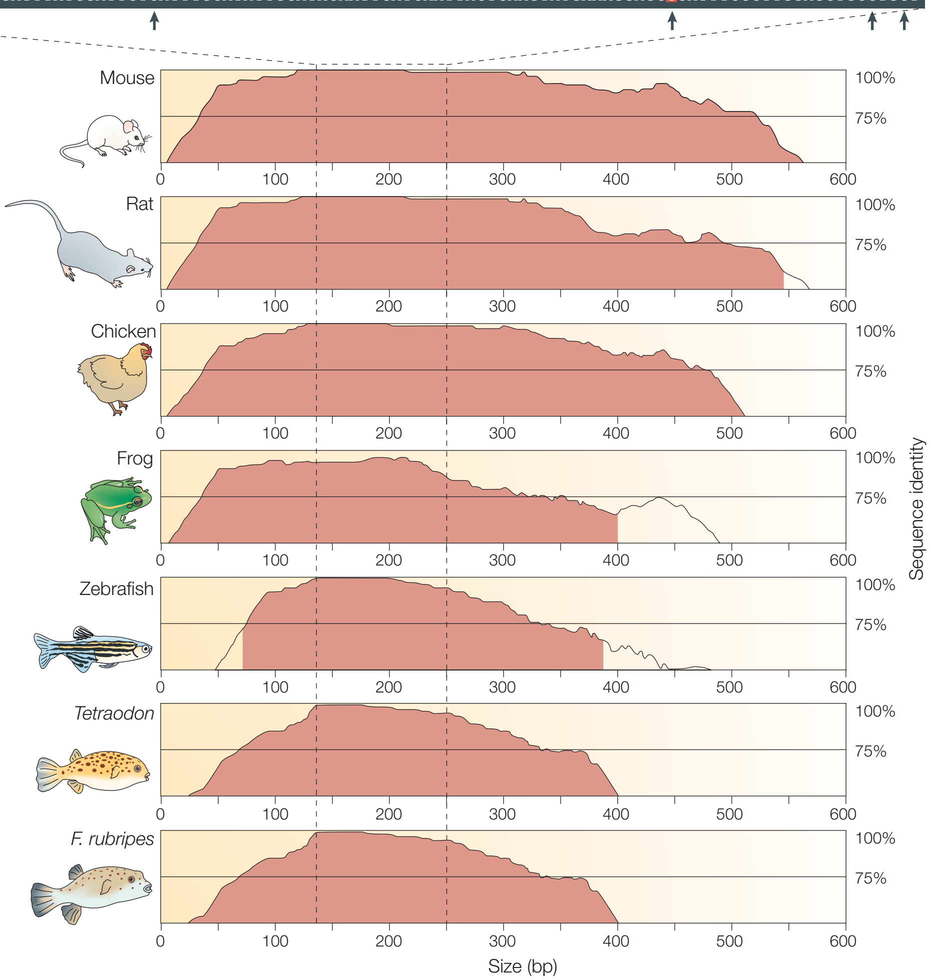
Mutations in these non-coding sequences that are conserved over long evolutionary periods can have an important role as a basis for human disease. To this end, it has been recently suggested that mutations in an enhancer that is conserved between human and *F. rubripes* causes a form of preaxial polydactyly, a common limb malformation in children. This enhancer regulates the topology of expression of sonic hedgehog (*Shh*) in limbs, from a distance of 1 million bp[21,32,33] (FIG. 2).

Enhancers conserved in distant vertebrates. The findings that mammals and fish share orthologous regulatory sequences that control the expression of orthologous genes should not have been completely surprising, given that they share many genes with similar tissue and temporal expression characteristics[17]. Why, then, is only a subset of the enhancers that are predicted to be shared between mammals and fish detected in human–fish alignments? One probable explanation is that simple nucleotide-alignment tools are inadequate for detecting many enhancers, and that tools that incorporate extra constraints, such as clustering[34,35], spacing[36] and patterns of evolution of TFBSs[37], might be required. Another possibility is that the enhancers that we do identify using human–fish comparisons represent a subset with uniquely rigorous structural and functional requirements. It is possible to imagine that these enhancers might have particular architectural constraints that prevent changes, such as inversions, insertions and deletions or nucleotide substitutions, that are usually tolerated by enhancers (BOX 2). This would preserve the sequence identity of these enhancers in distantly related genomes over a stretch of sequence that is large enough to be detected by nucleotide-alignment tools.

A recent study of an enhancer that regulates the expression of the homeobox *HOXC8* gene, which is conserved in humans and fish, provided evidence in favour of this hypothesis. An inversion of a single TFBS within this *HOXC8* enhancer, which contains multiple well-characterized TFBSs, results in the alteration of the spatial expression of a reporter gene that is driven by this enhancer[38]. Moreover, it seems that not only can the spatial arrangement and size of TFBSs within enhancers be conserved among distantly related species, but in

Box 2 | Extreme conservation in enhancers that are shared by human and fish

Human	GCTCACATAAATTAAGCTGTCATTCATAAGATTTATGGATTCTCATTTGCATATTGCATACAATTCATCAATTACTCAAGTATGAAAGGAGCACATTTCCCTTGGAGCTGCCTGCT
Mouse	GCTCACATAAATTAAGCTGTCATTCATAAGATTTATGGATTCTCATTTGCATATTGCATACAATTCATCAATTACTCAAGTATGAAAGGAGCACATTTCCCTTGGAGCTGCCTGCT
Rat	GCTCACATAAATTAAGCTGTCATTCATAAGATTTATGGATTCTCATTTGCATATTGCATACAATTCATCAATTACTCAAGTATGAAAGGAGCACATTTCCCTTGGAGCTGCCTGCT
Chicken	GCTCACATAAATTAAGCTGTCATTCATAAGATTTATGGATTCTCATTTGCATATTGCATACAATTCATCAATTACTCAAGTATGAAAGGAGCACATTTCCCTTGGAGCTGCATGCT
Frog	GCTCACATAAATTAAGCTATTATTCATAAGATTTATGGATTCTTATTTGCATATTGCATACAATTCATCAATTACTCAAGTATGAAAGGAGCACATTTCCCTTGGAGCTGCCTGAT
Zebrafish	GCTCACATAAATTAAGCTGTCATTCATAAGATTTATGGATTCTCATTTGCATATTGCATACAATTCATCAATTACTCAAGTATGAAAGGAGCACATTTCCCTTGGAGCTGCCTGCT
Tetraodon	GCTCACATAAATTAAGCTGTCATTCATAAGATTTATGGATTCTCATTTGCATATTGCATACAATTCATCAATTACTCAAGTATGAAAGGAGCGCATTTCCCTTGGAGCTGCCTGCT
F. rubripes	GCTCACATAAATTAAGCTGTCATTCATAAGATTTATGGATTCTCATTTGCATATTGCATACAATTCATCAATTACTCAAGTATGAAAGGAGCGCATTTCCCTTGGAGCTGCCTGCT



The finding that some *cis*-regulatory elements have almost identical sequences over hundreds of base pairs in species as far apart as humans and fish is surprising. Regulatory elements are generally composed of multiple transcription-factor binding sites (TFBSs) that are arranged in modules[18]. These modules are usually separated from each other by sequences, the length and identity of which are even more degenerate and flexible that the TFBSs themselves. As a result, regulatory sequences can tolerate small insertions, deletions or sequence substitutions. Regulatory variants such as these are an important source of phenotypic evolution[76]. So, in species as distant as humans and fish, it is more than reasonable to assume that in almost half a billion years, enough changes will have occurred within a given orthologous regulatory element so that even if these species still share this regulatory unit, their sequence will have diverged enough to render them 'invisible' to sequence-alignment tools.

Nevertheless, we can easily identify conserved enhancers between humans and fish. One of the reasons for this success is the astonishing degree of conservation that these sequences have retained over long periods. For example, a core enhancer in an intron in *DACH* is >98% identical for 350 bp in humans, mice and rats (see figure displaying 120 bp of the sequence alignment). Moreover, in the ~1 billion years of parallel evolutionary time that separates human, mouse, rat, chicken, frog and fish, only 6 substitutions occurred in a 120-bp fragment that corresponds to an enhancer[20], 4 of which occurred in the frog lineage alone, and none occurred in the mammalian lineage.

Arguably, as astonishing as their degree of conservation is the fact that these sequences correspond to enhancers. Why are these regulatory elements much more constrained than most other functional elements in the genome? Even if we picture a multi-modular enhancer, with overlapping modules that each have several TFBSs located on top of one another, it is still hard to imagine that this would justify such a degree of conservation over stretches of several hundred bases. It will be interesting to see what the functional investigation of these enhancers reveals, as to whether they have unique architectural features or whether they use similar mechanisms for transcription activation as other classical enhancers.

some cases, even the sequences of spacers that separate these TFBSs have remained untouched throughout long evolutionary periods. This extreme level of constraint on sequence variation was found in an enhancer that regulates the expression of *HOX* genes[39]. Therefore, it seems that constraints, that are as yet unknown, might be shaping the conservation of these non-coding sequences over extended evolutionary periods.

Clusters of conserved sequences. Another interesting aspect of human–fish conserved non-coding sequences is that they are predominantly found in clusters (FIG. 1). So, why are non-coding sequences that are conserved among evolutionarily distant species clustered around a subset of genes and absent from the regions that contain most genes? Most of the reported non-coding sequences that are conserved between humans and fish are found

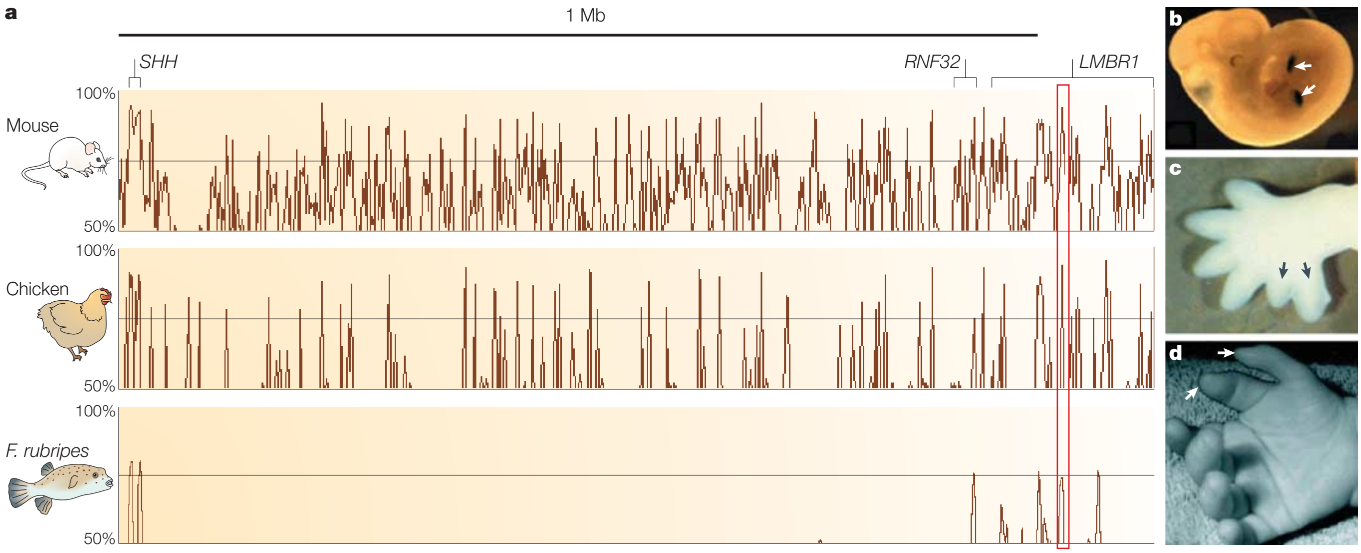


Figure 2 | **Sonic hedgehog expression in the limbs is regulated by an enhancer at a distance of 1 Mb. a** | Human–*Fugu rubripes* sequence comparisons, generated by VISTA, identify a conserved non-coding sequence in intron 5 of *LMBR1* (red box), which drives the expression of a reporter gene in a pattern that resembles the expression of sonic hedgehog (*SHH*) (arrows in **b**). Insertional mutagenesis in this region in mice results in preaxial polydactly (arrows in **c**). In humans, mutations in this enhancer are also associated with preaxial polydactyly (arrows in **d**). Adapted with permission from REF. 21. © (2003) Oxford University Press and REF. 32 (1999) Elsevier Science Ltd.

near genes the products of which have pivotal roles in embryonic development (TABLE 1). Such genes are known to be highly constrained, with a conserved pattern of expression among species across extreme evolutionary distances[40,41]. It is therefore plausible that some of the *cis*-regulatory sequences that control these genes are similarly constrained to precisely preserve the expression levels and patterns of genes that are crucial for basic vertebrate development. The activity of some of these enhancers that are involved in crucial steps in early development might need to be so tightly controlled such that even minute variations in activity (resulting from, say, a single nucleotide substitution) would be deleterious. However, it remains to be determined whether SNPs within these sequences occur at much lower frequencies than in other functional sequences of the genome. If this is the case, we might predict that sequence variants within these elements would markedly increase the likelihood of being associated with a phenotype. Recently, Haussler *et al.* have reported that the frequency of SNPs in sequences that are absolutely conserved between humans and rodents (100% identity, >200 bp), most of which are also conserved between humans and fish, is indeed several-fold lower than in other genomic sequences[42].

Limitations of distant comparisons. The unique features of distant species sequence comparisons highlighted above also point to some of the limitations that are inherent in their usefulness in annotating the human genome. The limited number of human–fish conserved non-coding elements, and their clustering around a few genes in the human genome, implies that most of these conserved sequences will contribute to the understanding of the regulation of only a subset of genes in the genome. Consequently, the analysis of genomes of species that are evolutionarily closer to humans than fish will become valuable for the identification of functional sequences otherwise missed by human–fish comparisons. For example, a recent analysis of *MEF2C* that compared human and chicken sequences identified an enhancer that regulates this gene's expression, which cannot be detected in human–fish sequence alignments[43].

Clearly, comparisons that use sequences of species from different evolutionary distances will bias the discovery of functional sequences to those that regulate biological features that are shared by the species being compared. Many non-coding sequences with crucially important roles will not be shared between human and *F. rubripes*, including sequences that were lost in one species, sequences that occurred in one lineage after the human and fish most recent common ancestor and sequences that have diverged beyond recognition. This highlights the crucial importance of choosing an appropriately distant model organism for the identification of *cis*-regulatory sequences. The recent availability of genomes at intermediate distances between fish and placental mammals, such as the frog *Xenopus tropicalis* and the chicken *Gallus gallus*, and the planned sequencing of the opossum *Monodelphis domestica*, will simplify the fine-tuning of the choice of species for comparative genomic analysis of any given gene.

Comparisons among primates

Comparisons between the genomes of closely related species, such as human and non-human primates, have been frequently dismissed as uninformative, owing to their inherent high sequence similarity. The

VISTA
A powerful tool for aligning the genome and visualizing the location of conserved sequences (see online links box).

few differences that have been detected between closely related species, however, hide within their folds biological insights that are not available from comparisons between species separated by a longer independent evolutionary history. In this section, we describe how comparisons among several primates have begun to reveal sequences that are conserved among primates, and that are in some instances missing in more distant species. We also discuss how comparisons between humans and their closest extant relative, chimpanzees, have allowed the identification of human-specific changes in protein-coding sequences.

Phylogenetic shadowing: to identify primate-specific conserved sequences. Two categories of regulatory element that are active in primates can potentially be identified through primate sequence comparisons, but are undetectable in comparisons with non-primates: regulatory elements that arose in the primate lineage and that are responsible for phenotypes unique to primates, and elements that, despite being derived from a common ancestral sequence and directing similar functions, have accumulated so many sequence changes that they show little sequence similarity between moderately distant species. Elegant examples of the latter elements have come from model organisms[44,45]. In a paradigmatic analysis of the *even-skipped* enhancer in *Drosophila melanogaster* and *Drosophila pseudoobscura*[44], 2 species for which the most recent common ancestor occurred 40–60 million years ago, Ludwig *et al.* clearly showed that, although this enhancer drives the same detailed expression pattern in both species, the underlying sequences are highly dissimilar. In other words, the sequence of the ancestral *even-skipped* enhancer has gradually changed in the two fly species, whereas the functional activity has not. The existence of the first type of regulatory elements, those that evolved *de novo* in primates, is still not proved, but it is reasonable to assume that the same sort of process that gives birth to new genes subsequent to gene duplication[46,47] can also give rise to novel regulatory elements.

Comparisons of several closely related primates would allow the identification of both types of enhancer. The paucity of sequence variation observed among primate sequences, however, requires a different approach from traditional pairwise comparative genomics. Recently, an approach, dubbed PHYLOGENETIC SHADOWING (BOX 3), was developed to reveal highly conserved sequences, which often correlate with functional regions[9]. The use of the phylogenetic shadowing approach was most clearly illustrated in the analysis of the regulatory sequence of *LPA*, a gene that contributes to heart disease in humans. Human *LPA* is one of a small set of genes that arose recently in the primate lineage and is consequently found in only a subset of primates[48,49]. Accordingly, intra-primate comparisons are the only comparative approach available to annotate *LPA*. As expected, sequence comparisons between the 5′ region of *LPA* in pairs of primate species revealed few sequence differences between non-functional regions and previously characterized functional regions. However, the collective sequence variation in the same region detected in 16 different primate genomes successfully revealed the location of numerous functional elements, including regulatory elements that were shown experimentally to be involved in this gene's expression[9]. Although these 'shadowing' studies included the sequence from ten or more primate species, modelling of the data indicated that sequences from as few as four non-human primate species, carefully chosen to include those least related to humans and to each other, can provide comparable resolution.

Table 1 | **Genes in proximity to highly conserved non-coding sequences***

Gene	Molecular function	Biological process	Reference
HOXB4	DNA-binding	Embryonic development	19
WNT1	Signal transducer	Embryonic development	79
SHH	Hydrolase and peptidase	Embryonic development	80,21
SCL (*TAL1*)	DNA-binding	Cell differentiation	29
SOX9	DNA-binding	Cell differentiation	81
DLL1	Protein-binding	Embryonic development	82
DLX1, -2, -5 and *-6*	DNA-binding	Embryonic development	24,31
HOXA1–13	DNA-binding	Embryonic development	83
HOXD cluster	DNA-binding	Embryonic development	84
DACH	Transcription factor	Embryonic development	20
NEUROG1	DNA-binding	Embryonic development	25
HOXC8	DNA-binding	Embryonic development	38
OTX2	DNA-binding	Embryonic development	28
CTGF	Growth-factor signalling	Cell growth/proliferation	26
PAX6	DNA-binding	Embryonic development	85
RUNX2	DNA-binding	Skeletal development	86

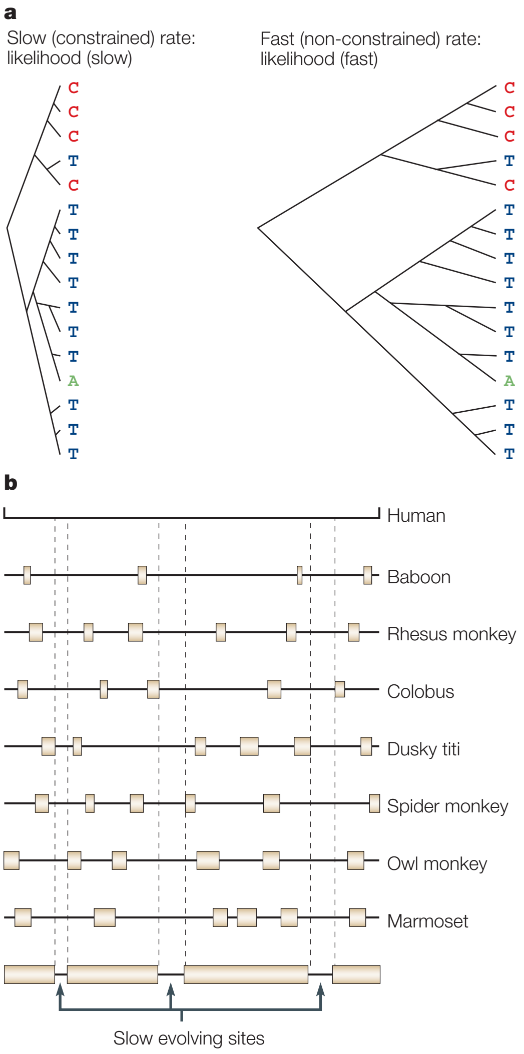
*Between mammals and fish. The molecular function and biological process of each gene were obtained from the Gene Ontology Consortium database (see online links box).

PHYLOGENETIC SHADOWING
An approach that combines comparisons of sequences from multiple, closely related species with a molecular phylogenetic model of sequence evolution to identify significantly conserved elements.

Box 3 | Phylogenetic shadowing

Phylogenetic shadowing analyses sequence variation in a multiple alignment to identify regions that accumulate variation at a slower rate. Each position in the multiple alignment is fitted to a phylogenetic model to calculate the likelihood that the position is evolving at a fast or a slow rate (**a**). Generally, positions with several sequence differences in multiple branches of the phylogenetic tree are more likely to be evolving at a fast rate, and in turn identify the least variable regions (**b**). The slowly evolving regions often correspond to functional sequences.

The use of highly similar sequences minimizes ambiguity in the computation of the multiple alignment. Moreover, the phylogenetic tree that relates the data is easy to infer and facilitates the comparative assembly of draft sequence from non-human primates to the reference human genome.


a
Slow (constrained) rate: likelihood (slow)
Fast (non-constrained) rate: likelihood (fast)
C C C T C T T T T T T T A T T T
C C C T C T T T T T T T A T T T
b
Human
Baboon
Rhesus monkey
Colobus
Dusky titi
Spider monkey
Owl monkey
Marmoset
Slow evolving sites

POSITIVE SELECTION
A sequence change in a species that results in increased fitness is subject to positive selection. As a consequence, the change normally becomes fixed, leading to adaptive evolution of that species.

Human–chimpanzee comparisons to identify genes that undergo adaptive evolution in humans. The extremely high degree of similarity between proteins from humans and chimpanzees prompted Wilson and King to suggest in a classic paper three decades ago that physiological differences between these two species would probably be explained by sequence changes not in genes but rather in regulatory elements[50]. Although this might still be the case, the increasing availability of sequences for humans and chimpanzees, and the development of sensitive computational tools for detecting POSITIVE SELECTION in protein-coding sequences[51,52], have allowed investigators to show that small sequence changes within the coding regions of genes also have an adaptive role in human evolution. Many of the genes that were first found to be under positive selection (BOX 4) are involved in the immune response[53] and in sexual reproduction[54–56]. These two classes of gene are known to be highly divergent among related species, an observation that is probably explained by the strong evolutionary constraints imposed by pathogen–host interactions and competition among sperm from different males to be the first to fuse with the egg[57]. More recently, genes that are believed to be involved in supporting an expanded brain[58], the ability to articulate sounds[59], the maintenance of genomic integrity[60], the development of masticatory muscles and the determination of jaw bone size[61] have been found to be undergoing adaptive evolution in humans, indicating that adaptive evolution has occurred for a wide variety of genes in the human genome.

The availability of genome sequences from humans, chimpanzees and mice recently allowed an unbiased genome-wide survey to identify positively selected genes in humans[62]. In that study, Clark *et al.* determined that ~9% of the exon-coding sequences, many of which could be putatively linked to physiological differences between humans and chimpanzees, have undergone adaptive evolution in the human lineage. These human–chimpanzee comparisons support the idea that small incremental changes in protein sequences might underlie at least part of the phenotypic diversity between humans and their closest phylogenetic relatives. Although the identification of adaptively evolving genes does not prove that the positively selected amino-acid replacements altered the function of the resulting protein, we might predict that they frequently occur at functionally important sites, indicating an obvious starting point for functional analyses[63,64].

Adaptive evolution in non-coding sequences. Can we expect that the availability of the draft sequence of the chimpanzee genome will support an advance in the identification of those adaptive changes in non-coding functional elements that Wilson and King predicted to be important in human evolution? Unlike in coding sequences, we do not know how to differentiate neutral from functional changes in non-coding sequences. This stems in part from our incomplete understanding of functional non-coding elements of the genome. Of those, the best understood are gene regulatory elements to which transcription factors bind, such as promoters, enhancers and locus control regions. The degeneracy of TFBSs[65] makes it difficult to use these sequences as a regulatory equivalent of the genetic code. For these reasons, it will be difficult with sequence comparisons alone to identify adaptive changes in gene regulatory elements that distinguish humans from chimpanzees: extensive experimental validation will be required[66,67]. Although the sequence of the chimpanzee genome will probably provide limited insights into the regulatory origin of human traits, it will certainly reveal small- and large-scale genomic rearrangements that contribute to the DNA differences between chimpanzees and humans. Studies that compare the sequence of human chromosome 21 to the corresponding chimpanzee chromosome[68,69] have already shown that insertions and deletions occurred frequently during primate evolution and that they are an important component of genome differences between humans and chimpanzees. Duplications and deletions of gene- or regulatory element-containing regions will probably contribute in some cases to differences in gene-expression levels between chimpanzees and humans.

Comparative genomics within a single species. The successful application of phylogenetic shadowing raises a more compelling question: is it possible to use the sequence polymorphisms found in human populations to annotate the human genome? Such an approach would circumvent many of the problems connected with using model organisms, such as their divergence in physiology and sequence. Although the low degree of polymorphism in humans makes such studies challenging at first glance, the marked acceleration of human-genome resequencing that will probably occur in the future might make this strategy feasible. Preliminary studies that involve the resequencing and analysis of genomic intervals from many *Ciona intestinalis* individuals — an organism with high polymorphism frequency — have indicated that intra-species sequence comparisons can be successfully used to identify gene regulatory elements and exons (D.B. and E.M.R., unpublished observations).

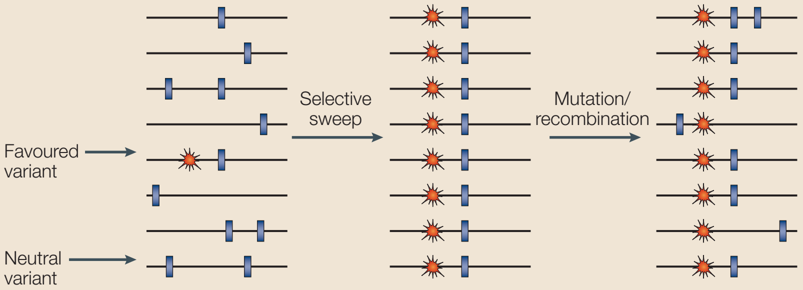
The complexity of human-population dynamics and haplotype structure make it difficult to estimate the number of individuals required to apply this approach to humans. Nonetheless, a simple extrapolation of the data from Yu *et al.*[70] indicates that sequencing fewer than 1,000 individuals would yield 0.3 SNPs per site, a number that is comparable to that used in previous studies[9]. Although it is probable that the number of individuals required for intra-human comparative genomics will turn out to be greater than this estimate, the predicted increase in human genome resequencing[71] should prove that such comparisons are fruitful for the identification of functional sequences that are shared with other species as well as those that are unique to our species.

Future outlook

Genome comparisons at the extremes of the evolutionary range have important advantages and limitations. Distant species comparisons help to reveal a subset of extremely conserved non-coding elements shared between all vertebrates that are associated with an easily accessed function. However, these elements are not distributed uniformly across the genome but tend to cluster around a small subset of genes, thereby limiting their general application. Of particular interest is that they tend to be clustered around DNA- and RNA-binding proteins that are expressed early in development, indicating that this category of gene regulatory elements might have fundamental constraints for the viability of the developing organism. At the other end of the scale, primate comparisons allow the identification of primate-specific functional elements, which are unavailable from more distant species comparisons. However, they lack the statistical power to identify short stretches of conservation, such as those that correspond to individual TFBSs. The annotation of the human genome to this level of resolution will probably require the simultaneous analysis of several mammalian species that are separated by similar, intermediate evolutionary distances[8,72] (BOX 1).

The ability of multiple species comparisons to reveal fine-scale features, such as individual TFBSs, has been elegantly shown in an analysis of four related yeast species by Kellis *et al.*[7]. Cooper *et al.* estimated that the

Box 4 | How do we identify adaptively evolving genes?



Comparing the sequence of a gene between two related species reveals the nucleotide changes that have occurred since their last common ancestor. An important problem in the search for genes that undergo adaptive evolution is distinguishing the changes that have been positively selected from the neutral ones. The redundancy of the genetic code comes to the rescue, allowing us to distinguish two types of change in exon-coding sequences: synonymous substitutions (those that leave the encoded amino acid unchanged and are therefore selectively neutral; d_S), and non-synonymous substitutions (those that result in a change in amino-acid sequence and are subject to selection; d_N). The ratio of non-synonymous to synonymous substitutions, d_N/d_S, indicates the type of selection that a gene is subject to. An excess of non-synonymous substitutions in pairwise sequence comparisons ($d_N/d_S >1$) indicates that 1 of the 2 sequences is undergoing positive selection. To determine in which lineage positive selection occurred, however, the sequence of a third species is needed.

Increases in d_N/d_S ratios can also result from a decrease in effective population size or a relaxation of selective constraints. Evidence of a 'selective sweep' is often sought to discriminate between these two possibilities and positive selection[77]. If a change in a gene is positively selected, linked neutral variation 'hitchhikes' along with the selected site[78]. After the selected change has 'swept' through a species, variation begins to build up again around it, resulting in an excess of rare polymorphisms as a signature of the sweep. Selective sweeps can only be used to reveal recent episodes of positive selection. With time, new mutations will occur and recombination will break the linkage between selected and neutral variants, removing the evidence of the sweep.

sequence of 10–15 mammalian species would be required to achieve a similar level of resolution in the much more complex human genome[73]. However, the much higher ratio of non-functional to functional sequences in the human relative to the yeast genome makes it unclear whether this type of resolution will be achievable in complex genomes. Projects such as the recently launched Encode (see online links box), which is designed to exploit comparative genomics at a great range of evolutionary distances, and various computational and experimental approaches to exhaustively annotate defined regions that comprise 1% of the human genome, will clearly assist in the development of a roadmap for future approaches to the fine annotation of the human genome.

Despite the power of experimental biology, it is probable that many functionally conserved non-coding elements will remain 'experimentally invisible', owing to either the inadequacies of our biological assays, which are mostly limited to certain aspects of genomic function such as transcriptional regulation, or to the small impact that each of these elements might individually have. Consequently, after we have completed the characterization of the most strongly conserved elements in the genome — those with readily assignable functions — how can we functionally annotate the remaining elements? Barring the development of ultrasensitive biological assays in the near future, one of the approaches to deal with these sequences might not be to define their exact function but rather to assess their relative functional importance to the organism. With the availability of the sequence of many species, we could start to assign quantitative scores to each of these elements reflecting the likelihood that they have been functionally retained in the genomes of humans and other species by natural selection. Ultimately, with the cost of sequencing significantly reduced, we should be able to resequence human genomes on a large scale. Given a large enough sample and exhaustive human phenotypic information, we might be able to develop a quantitative matrix that assigns morbidity scores to each base of the human genome on the basis of the degree of variation observed at that position. An initial step in this daunting quest is to turn our focus to those low hanging fruits that nature has already shown us to be important by keeping them untouched in our genome.

1. Homer. *The Odissey* Ch. 12 (Signet Classic, New York, 1999).
2. Nobrega, M. A. & Pennacchio, L. A. Comparative genomic analysis as a tool for biological discovery. *J. Physiol.* **554**, 31–39 (2004).
3. Pennacchio, L. A. & Rubin, E. M. Comparative genomic tools and databases: providing insights into the human genome. *J. Clin. Invest.* **111**, 1099–1106 (2003).
4. Frazer, K. A. *et al.* Evolutionarily conserved sequences on human chromosome 21. *Genome Res.* **11**, 1651–1659 (2001).
5. Loots, G. G. *et al.* Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288**, 136–140 (2000).
6. Pennacchio, L. A. *et al.* An apolipoprotein influencing triglycerides in humans and mice revealed by comparative sequencing. *Science* **294**, 169–173 (2001).
7. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E. S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**, 241–254 (2003).
Paradigmatic example of the power of comparisons of multiple, related genomes to identify functional sequence in a genome.
8. Thomas, J. W. *et al.* Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424**, 788–793 (2003).
9. Boffelli, D. *et al.* Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299**, 1391–1394 (2003).
The first paper to describe the use of comparisons of multiple, closely related primates to identify primate-specific conserved sequences.
10. Hardison, R. C. *et al.* Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Res.* **13**, 13–26 (2003).
11. Hardison, R. C. Comparative genomics. *PLoS Biol.* **1**, E58 (2003).
12. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
13. Schwartz, S. *et al.* Human–mouse alignments with BLASTZ. *Genome Res.* **13**, 103–107 (2003).
14. Pennacchio, L. A., Baroukh, N. & Rubin, E. M. in *Symposia on Quantitative Biology: The Genome of* Homo sapiens (Cold Spring Harbor Press, Cold Spring Harbor, in the press).
15. Elnitski, L. *et al.* Distinguishing regulatory DNA from neutral sites. *Genome Res.* **13**, 64–72 (2003).
16. Brenner, S. *et al.* Characterization of the pufferfish (*Fugu*) genome as a compact model vertebrate genome. *Nature* **366**, 265–268 (1993).
17. Aparicio, S. *et al.* Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297**, 1301–1310 (2002).
18. Arnone, M. I. & Davidson, E. H. The hardwiring of development: organization and function of genomic regulatory systems. *Development* **124**, 1851–1864 (1997).
19. Aparicio, S. *et al.* Detecting conserved regulatory elements with the model genome of the Japanese puffer fish, *Fugu rubripes*. *Proc. Natl Acad. Sci. USA* **92**, 1684–1688 (1995).
Demonstrates that human–*F. rubripes* comparisons detect conserved non-coding sequences that, once tested in *in vivo* assays, correspond to enhancers.
20. Nobrega, M. A., Ovcharenko, I., Afzal, V. & Rubin, E. M. Scanning human gene deserts for long-range enhancers. *Science* **302**, 413 (2003).
21. Lettice, L. A. *et al.* A long-range *Shh* enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum. Mol. Genet.* **12**, 1725–1735 (2003).
Demonstrates that sequence variation in *cis*-regulatory elements at near-megabase distances can result in phenotypic variation.
22. Kleinjan, D. J. & van Heyningen, V. Position effect in human genetic disease. *Hum. Mol. Genet.* **7**, 1611–1618 (1998).
23. de Kok, Y. J. *et al.* Identification of a hot spot for microdeletions in patients with X-linked deafness type 3 (DFN3) 900 kb proximal to the DFN3 gene *POU3F4*. *Hum. Mol. Genet.* **5**, 1229–1235 (1996).
24. Zerucha, T. *et al.* A highly conserved enhancer in the *Dlx5/Dlx6* intergenic region is the site of cross-regulatory interactions between *Dlx* genes in the embryonic forebrain. *J. Neurosci.* **20**, 709–721 (2000).
25. Blader, P., Plessy, C. & Strahle, U. Multiple regulatory elements with spatially and temporally distinct activities control neurogenin1 expression in primary neurons of the zebrafish embryo. *Mech. Dev.* **120**, 211–218 (2003).
26. Dickmeis, T. *et al.* Expression profiling and comparative genomics identify a conserved regulatory region controlling midline expression in the zebrafish embryo. *Genome Res.* **14**, 228–238 (2004).
27. Goode, D. K., Snell, P. K. & Elgar, G. K. Comparative analysis of vertebrate *Shh* genes identifies novel conserved non-coding sequence. *Mamm. Genome* **14**, 192–201 (2003).
28. Kimura-Yoshida, C. *et al.* Characterization of the pufferfish *Otx2 cis*-regulators reveals evolutionarily conserved genetic mechanisms for vertebrate head specification. *Development* **131**, 57–71 (2004).
29. Barton, L. M. *et al.* Regulation of the stem cell leukemia (*SCL*) gene: a tale of two fishes. *Proc. Natl Acad. Sci. USA* **98**, 6747–6752 (2001).
30. Lien, C. L., McAnally, J., Richardson, J. A. & Olson, E. N. Cardiac-specific activity of an *Nkx2-5* enhancer requires an evolutionarily conserved Smad binding site. *Dev. Biol.* **244**, 257–266 (2002).
31. Ghanem, N. *et al.* Regulatory roles of conserved intergenic domains in vertebrate *Dlx* bigene clusters. *Genome Res.* **13**, 533–543 (2003).
32. Sharpe, J. *et al.* Identification of *Sonic hedgehog* as a candidate gene responsible for the polydactylous mouse mutant *Sasquatch*. *Curr. Biol.* **9**, 97–100 (1999).
33. Lettice, L. A. *et al.* Disruption of a long-range *cis*-acting regulator for *Shh* causes preaxial polydactyly. *Proc. Natl Acad. Sci. USA* **99**, 7548–7553 (2002).
34. Berman, B. P. *et al.* Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proc. Natl Acad. Sci. USA* **99**, 757–762 (2002).
35. Markstein, M., Markstein, P., Markstein, V. & Levine, M. S. Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the *Drosophila* embryo. *Proc. Natl Acad. Sci. USA* **99**, 763–768 (2002).
36. Chiang, D. Y., Moses, A. M., Kellis, M., Lander, E. S. & Eisen, M. B. Phylogenetically and spatially conserved word pairs associated with gene-expression changes in yeasts. *Genome Biol.* **4**, R43 (2003).
37. Moses, A. M., Chiang, D. Y., Kellis, M., Lander, E. S. & Eisen, M. B. Position specific variation in the rate of evolution in transcription factor binding sites. *BMC Evol. Biol.* **3**, 19 (2003).
38. Anand, S. *et al.* Divergence of *Hoxc8* early enhancer parallels diverged axial morphologies between mammals and fishes. *Proc. Natl Acad. Sci. USA* **100**, 15666–15669 (2003).
39. Mainguy, G. *et al.* A position-dependent organisation of retinoid response elements is conserved in the vertebrate *Hox* clusters. *Trends Genet.* **19**, 476–479 (2003).
40. Erwin, D. H. & Davidson, E. H. The last common bilaterian ancestor. *Development* **129**, 3021–3032 (2002).
One of the many insightful studies by this group that characterizes genetic regulatory networks, aspects of which are shared by all bilaterians, in contrast to other aspects that probably evolved later, in subgroups of species.
41. Davidson, E. H. *et al.* A genomic regulatory network for development. *Science* **295**, 1669–1678 (2002).
42. Bejerano, G. *et al.* Ultra-conserved elements in the human genome. *Science* 6 May 2004 (doi:10.1126/science.1098119).
Seminal study first reporting the characterization of ultra-conserved elements in mammalian genomes.
43. Dodou, E., Xu, S. M. & Black, B. L. *mef2c* is activated directly by myogenic basic helix-loop-helix proteins during skeletal muscle development *in vivo*. *Mech. Dev.* **120**, 1021–1032 (2003).
44. Ludwig, M. Z., Bergman, C., Patel, N. H. & Kreitman, M. Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature* **403**, 564–567 (2000).
First convincing demonstration of the role of balancing selection in maintaining an invariant function in enhancers with diverging sequence.
45. Takahashi, H., Mitani, Y., Satoh, G. & Satoh, N. Evolutionary alterations of the minimal promoter for notochord-specific *Brachyury* expression in ascidian embryos. *Development* **126**, 3725–3734 (1999).

46. Lynch, M. & Conery, J. S. The evolutionary fate and consequences of duplicate genes. *Science* **290**, 1151–1155 (2000).

47. Johnson, M. E. *et al.* Positive selection of a gene family during the emergence of humans and African apes. *Nature* **413**, 514–519 (2001).

48. Lawn, R. M. *et al.* The recurring evolution of lipoprotein(a). Insights from cloning of hedgehog apolipoprotein(a). *J. Biol. Chem.* **270**, 24004–24009 (1995).

49. Boffelli, D., Cheng, J. F. & Rubin, E. M. Convergent evolution in primates and an insectivore. *Genomics* **83**, 19–23 (2004).

50. King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188**, 107–116 (1975).

51. Yang, Z. & Nielsen, R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* **17**, 32–43 (2000).

52. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**, 555–556 (1997).

53. Hughes, A. L. & Yeager, M. Natural selection at major histocompatibility complex loci of vertebrates. *Annu. Rev. Genet.* **32**, 415–435 (1998).

54. Swanson, W. J., Yang, Z., Wolfner, M. F. & Aquadro, C. F. Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. *Proc. Natl Acad. Sci. USA* **98**, 2509–2514 (2001).

55. Swanson, W. J. & Vacquier, V. D. The rapid evolution of reproductive proteins. *Nature Rev. Genet.* **3**, 137–144 (2002).

56. Wyckoff, G. J., Wang, W. & Wu, C. I. Rapid evolution of male reproductive genes in the descent of man. *Nature* **403**, 304–309 (2000).

57. Clark, A. G., Begun, D. J. & Prout, T. Female x male interactions in *Drosophila* sperm competition. *Science* **283**, 217–220 (1999).

58. Goldberg, A. *et al.* Adaptive evolution of cytochrome c oxidase subunit VIII in anthropoid primates. *Proc. Natl Acad. Sci. USA* **100**, 5873–5878 (2003).

59. Enard, W. *et al.* Molecular evolution of *FOXP2*, a gene involved in speech and language. *Nature* **418**, 869–872 (2002).
Elegant identification of a gene suspected to be involved in the development of speech undergoing positive selection in the human lineage.

60. Huttley, G. A. *et al.* Adaptive evolution of the tumour suppressor *BRCA1* in humans and chimpanzees. Australian Breast Cancer Family Study. *Nature Genet.* **25**, 410–413 (2000).

61. Stedman, H. H. *et al.* Myosin gene mutation correlates with anatomical changes in the human lineage. *Nature* **428**, 415–418 (2004).

62. Clark, A. G. *et al.* Inferring nonneutral evolution from human–chimp–mouse orthologous gene trios. *Science* **302**, 1960–1963 (2003).

63. Zhang, J., Zhang, Y. P. & Rosenberg, H. F. Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. *Nature Genet.* **30**, 411–415 (2002).

64. Fleming, M. A., Potter, J. D., Ramirez, C. J., Ostrander, G. K. & Ostrander, E. A. Understanding missense mutations in the *BRCA1* gene: an evolutionary approach. *Proc. Natl Acad. Sci. USA* **100**, 1151–1156 (2003).

65. Wasserman, W. W. & Sandelin, A. Applied bioinformatics for the identification of regulatory elements. *Nature Rev. Genet.* **5**, 276–287 (2004).

66. Gumucio, D. L. *et al.* Differential phylogenetic footprinting as a means to identify base changes responsible for recruitment of the anthropoid γ-gene to a fetal expression pattern. *J. Biol. Chem.* **269**, 15371–15380 (1994).

67. Rockman, M. V., Hahn, M. W., Soranzo, N., Goldstein, D. B. & Wray, G. A. Positive selection on a human-specific transcription factor binding site regulating *IL4* expression. *Curr. Biol.* **13**, 2118–2123 (2003).

68. Frazer, K. A. *et al.* Genomic DNA insertions and deletions occur frequently between humans and nonhuman primates. *Genome Res.* **13**, 341–346 (2003).

69. Locke, D. P. *et al.* Large-scale variation among human and great ape genomes determined by array comparative genomic hybridization. *Genome Res.* **13**, 347–357 (2003).

70. Yu, N. *et al.* Larger genetic differences within Africans than between Africans and Eurasians. *Genetics* **161**, 269–274 (2002).

71. Collins, F. S., Green, E. D., Guttmacher, A. E. & Guyer, M. S. A vision for the future of genomics research. *Nature* **422**, 835–847 (2003).

72. Dermitzakis, E. T. *et al.* Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs). *Science* **302**, 1033–1035 (2003).

73. Cooper, G. M. *et al.* Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes. *Genome Res.* **13**, 813–820 (2003).

74. Margulies, E. H., Blanchette, M., NISC Comparative Sequencing Program, Haussler, D. & Green, E. D. Identification and characterization of multi-species conserved sequences. *Genome Res.* **13**, 2507–2518 (2003).

75. Frazer, K. A. *et al.* Noncoding sequences conserved in a limited number of mammals in the *SIM2* interval are frequently functional. *Genome Res.* **14**, 367–372 (2004).

76. Carroll, S. B. Endless forms: the evolution of gene regulation and morphological diversity. *Cell* **101**, 577–580 (2000).

77. Fay, J. C., Wyckoff, G. J. & Wu, C. I. Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. *Nature* **415**, 1024–1026 (2002).

78. Smith, J. M. & Haigh, J. The hitch-hiking effect of a favourable gene. *Genet. Res.* **23**, 23–35 (1974).

79. Gellner, K. & Brenner, S. Analysis of 148 kb of genomic DNA around the *wnt1* locus of *Fugu rubripes*. *Genome Res.* **9**, 251–258 (1999).

80. Muller, F. *et al.* Intronic enhancers control expression of zebrafish *sonic hedgehog* in floor plate and notochord. *Development* **126**, 2103–2116 (1999).

81. Bagheri-Fam, S., Ferraz, C., Demaille, J., Scherer, G. & Pfeifer, D. Comparative genomics of the *SOX9* region in human and *Fugu rubripes*: conservation of short regulatory sequence elements within large intergenic regions. *Genomics* **78**, 73–82 (2001).

82. Hans, S. & Campos-Ortega, J. A. On the organisation of the regulatory region of the zebrafish δD gene. *Development* **129**, 4773–4784 (2002).

83. Santini, S., Boore, J. L. & Meyer, A. Evolutionary conservation of regulatory elements in vertebrate *Hox* gene clusters. *Genome Res.* **13**, 1111–1122 (2003).

84. Spitz, F., Gonzalez, F. & Duboule, D. A global control region defines a chromosomal regulatory landscape containing the *HoxD* cluster. *Cell* **113**, 405–417 (2003).
One of the most elegant examples of the application of distant vertebrate sequence comparisons aiding the sifting of large genomic intervals for functional sequences.

85. Griffin, C., Kleinjan, D. A., Doe, B. & van Heyningen, V. New 3′ elements control *Pax6* expression in the developing pretectum, neural retina and olfactory region. *Mech. Dev.* **112**, 89–100 (2002).

86. Eggers, J. H., Stock, M., Fliegauf, M., Vonderstrass, B. & Otto, F. Genomic characterization of the *RUNX2* gene of *Fugu rubripes*. *Gene* **291**, 159–167 (2002).

Acknowledgements

We would like to thank Len Pennacchio and members of the Rubin laboratory for useful discussions. Research was conducted at the E. O. Lawrence Berkeley National Laboratory, with support by a grant from the Programs for Genomic Application, NHLBI, and performed under a Department of Energy contract, University of California, USA.

Competing interests statement

The authors declare that they have no competing financial interests.

Online links

DATABASES

The following terms in this article are linked online to:

Entrez: http://www.ncbi.nlm.nih.gov/Entrez

DACH | *even skipped* | *Hoxb4* | *LMBR1* | *LPA* | *MEF2C* | *Shh*

FURTHER INFORMATION

Encode: www.genome.gov/encode

Gene Ontology Consortium database: http://www.geneontology.org

SHADOWER: http://bonaire.lbl.gov/newshadower

VISTA Tools: http://gsd.lbl.gov/vista/index.shtml

Access to this links box is available online.